

Statistical Lip Geometry Estimation for Lip Reading

Alin Chițu[#], Léon J.M. Rothkrantz^{**}

[#]Delft University of Technology, The Netherlands

^{*}Netherlands Defence Academy, The Netherlands

l.j.m.rothkrantz@tudelft.nl

Abstract—Automatic lip reading can be used to enhance automatic speech recognition in noisy environments, as an interface for hearing impaired persons, security applications and speech recovery from mute or deteriorated films. Several techniques for modelling lip movements have been proposed. In this paper we present a method to describe the shape of the mouth based solely on a statistical interpretation of the distribution of the pixels that lie on the lips and the performance of lip reading systems train based on the features vectors defined using this method. This method is special because it does not rely on a priori defined model of the lips nor try to detect any special features of the mouth (i.e. it does not search for key points or other low level features). The method was first proposed by Wojdel and Rothkrantz in [1, 2]. In this paper we give a description of the algorithm and then we report the performance of the lip reader built based on this method. It also presents a method for the detection and the description of the teeth, tongue and cavity. These elements are very important for lip reading and can be used as extra features in any other setting, irrespective to the main feature extraction technique.

Keywords—Lip Reading; Lip Geometry; Face Analysis; Automatic Viseme Recognition

I. INTRODUCTION

Lip reading was thought for many years to be specifically used by hearing impaired persons. Therefore, lip reading was considered as a possible application in abnormal situations. Extensive lip reading research was primarily done in order to improve the teaching methodology for hearing impaired persons and to increase their chances for integration in society. Later, research done on human perception, more exactly speech perception proved that lip reading is actively employed in different degrees by all humans irrespective with their hearing capacity. The most well know study in this respect was performed by Harry McGurk and John MacDonald in 1976 [3]. In their experiment the two researchers tried to understand the perception of speech by children. Their findings, now known as the McGurk effect [3], was that if a person is presented a video sequence with a certain utterance (e.g. in their experiments utterance 'ga'), but in the same time the acoustics present a different utterance (e.g. in their experiments the sound 'ba'), in a large majority of cases the person will perceive a third utterance (i.e. in this case 'da'). Subsequent experiments showed that this is true as well for longer utterances and that it is not a particularity of the visual and aural senses but also as true for other perception functions. Therefore, lip reading is a part of our multi-sensory speech perception process and could be better named visual speech recognition. Being an evolutionary acquired capacity, same as speech perception, some scientists consider the lip reading's neural mechanism the one that enables humans to achieve high literacy skills with relative ease.

Another source of confusion is the "lip" word, because it implies that lips are the only part of the speaker's face that transmits visual information about what is being said. The

teeth, the tongue and the cavity were shown to be of great importance for lip reading by humans [4]. Also other face elements were shown to be important during face to face communication; however, their exact influence is not completely elucidated. During experiments in which a gaze tracker was used to track the speaker's areas of attention during communication, it was found that humans focus on four major areas: the mouth, the eyes and the centre of the face depending on the task and the noise level [5]. In normal situations the listener scans the mouth and the other areas with relatively equal amounts of time. However, when the background noise increases, the centre of the face becomes the central point of attention. Most probably the peripheral vision becomes extremely active in these situations. When the task was to infer the emotional load of the interlocutor, the listener's gaze started to shift towards the eyes, since they convey more emotional related information. It is well accepted that the human lip readers make great use of the context in which the interaction takes place. This can be one of the reasons why the human listener scans the entire face during the interaction. In Hilder [6] the authors found that when a human lip reader was presented with appearance information, compared with only mouth shapes, his performance increased considerably from 42.9% to 71.6%.

We should realise that during face to face interaction a human engages in a complex process which involves various channels of information corresponding to our senses. In this way the speaker builds up the context using both verbal and non-verbal cues such as body gesture, facial expressions, prosody, and other physiological manifestations. Other information about the settings in which the communication takes place is used as well as the knowledge accumulated in time through experience. Man is a multi-modal, multi-sensor, multi-media fusing machine.

The rest of the paper is organized as follows: in section 2 we present an overview of related work, section 3 starts by introducing the details of the algorithm as well as giving some examples of image filters that could be used in the pre-processing of the input image. Thereafter, we give some methods for making the algorithm more robust for instance by restricting the ROI and by refining the output of the filter through an outlier detection method. Section 4 presents the results of experiments. After giving some validating examples for the suitability of the method, we present the performance of the lip reader built based on this method. We end this paper with our conclusions in section 5.

II. RELATED WORK

It is more than three decades since the first experiments in automatic lip reading emerged from the scientific community. However, only until the 90s and more sustained the second half of the 90s, the subject started to become viable. Even

today it still lags speech recognition by some decades. Until recent years the most impeding factor was the computational power of the computers. Nowadays it is difficult to find the most suitable visual features that capture the information related with what is being spoken and the hard problem of accurately detecting and tracking the facial elements that convey speech related information. The automatic and robust detection and tracking of face elements is still not entirely achieved by the current technology. As in similar visual pattern recognition applications, the two monsters "illumination variations" and "occlusions" are still alive and menacing.

The task of isolated letters was among the first analyses by Petajan et al. [7] back in 1998. The authors reported that the correct recognition is close to 90%. However, based on the AVletters data corpus, Matthews et al. [8] reported only a 50% recognition rate. Li et al. [9] report a perfect recognition 100% on the same task, but two years later in Li et al. [10] only 90% recognition. The second most popular task is digit recognition either in isolation or in connected strings. Based on the TULIPS1 data corpus, which only contains the first four digits, Luetin et al. [11] and Luetin and Thacker [12] reported 83.3% and 88.5% recognition rates, respectively. Arsic and Thiran [13] reported the same data corpus 81.25% and 89.6% depending on the feature extraction method. Other experiments with the digit recognition task are: Potamianos et al. [14] reported 95.7%, Dupont and Luetin [15] reported 59.7%, Wojdel reported in his thesis [16] 91.1% correct recognition and 81.1% accuracy, Potamianos et al. [17] reported 63% and Perez et al. [18] 47%. Lucey and Potamianos [19] reported 74.6% recognition rate for the isolated digits task. Potamianos et al. [14] report 64.5% recognition rate for the connected letter task. For the isolated word task Nefian et al. [20] reported 66.9%, Zhang et al. [21] reported 42%, Kumar et al. [22] reported 42.3%. We can conclude that there is still a large variation in the performances obtained, and there is still no convergence visible since the newer studies do not necessarily show an increase in accuracy. This is, in our opinion, clearly a sign of the immaturity of the lip reading domain. Also, as can be observed in the listing above, there are yet no results of experiments with continuous speech. Patamianos et al. [17] report an extremely low result on the continuous speech task, namely 12% an extremely low result. The lip reading domain is still young and there are many limiting factors that need to be conquered. Nevertheless, as shown in many studies, lip reading can be successfully used in conjunction with speech for an enhanced speech recognition system.

III. MODEL

Our model uses an image filter to compute for each pixel the probability to be part of the lips. The result of the filter is then statistically interpreted in order to describe the shape of the mouth. The analysis of the shape of the mouth is independent of the choice of the image filter; however, the performance of the algorithm is strongly linked with the capacity of the filter to exactly identify the pixels that lie on the lips. An overview of the algorithm is depicted in Figure 1. The algorithm evolves a processing pipeline and contains the following steps: extraction of the next frame, ROI detection or tracking, application of an image filter to describe the pixels, analysis of the filter result and refinement, and finally analysis of the refined filter result and computation of the output

features. The following sections present in detail all steps described.

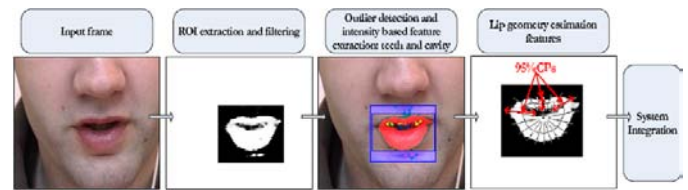


Fig. 1 The process of estimating the lip geometry shown on real example. From left to right we have the input image, the result of the ROI detection and image filter procedure, the outlier removal and the lips, cavity and teeth detection and finally the geometric features shown in polar co-ordinates.

A. Defining the Region of Interest

As the first step of the processing pipeline, we have to locate the face and then the mouth of the speaker. The reduction of the searching area (i.e. ROI) removes unnecessary parts from the image which is very important for at least two reasons: first the processing time is greatly reduced and second many possible unwanted artefacts can be avoided. For these we use the Viola-Jones algorithm for object detection [23]. This classifier uses a method to detect the most representative Haar like features using a learning algorithm based on AdaBoost. It combines a set of weak classifiers using a 'cascading' approach which corroborated with a fast method for computing the Haar-like features, allows high speed and very low false-negative rates. In order to increase the reliability of the ROI extraction process we used a combination of detection and tracking steps. Hence, in a first step the mouth of the speaker is detected using the Viola-Jones detector, and in the subsequent steps the mouth is tracked using an object tracking algorithm which is adapted using the last detected ROI. The object tracking algorithm uses a Gaussian Mixture Model to model the colour distribution of the object and of the background and a deformable template to optimally fit the tracked object.

B. Lips Segmentation

The next step in the process is to compute each pixel the probability that it belongs to the lips. Fortunately, because the input image contains only the mouth area and the lips usually have a distinct colouring, we can extract the lip's pixels without the need for complicated additional object recognition techniques. The lip selective filter is not fixed to any pre-chosen image filter and therefore any available method can be used. The only requirement is that the filter returns the result in probabilistic terms, namely, each pixel should be given a value between 0 and 1, 1 meaning maximum confidence that the given pixel belongs to the lips. It is very important to choose the most relevant colour encoding system. We tested several image filters based on different colour systems. As with all image segmentation techniques the illumination conditions and the quality of the recorded video sequences greatly influences the end result. A parabolic shaped filter is very simple to implement and from the point of view of computing power requirements very attractive. Unfortunately, during our experiments we found that in many cases, if the illumination of the face is not perfect, the hue component itself is not sufficient for proper lip selection. Even combining in a cascade the results from parabolic thresholding in different colour spaces did not yield sufficiently robust filters. On the other hand, we found that training a simple feed-forward neural network was performing much better for our data corpus. The network that was used has only 5 neurons in

a hidden layer and one output neuron. The network was trained using RGB values from the pixels in the image. This filter achieved extremely accurate results. The results obtained with several filters are shown in Figure 2, where it is visible that the neural network outperforms all the other approaches. In all situations the same area was used for training, namely the area covering the lower lip. One problem of the hue based filter is caused by the fact that the hue is not well defined for very bright reflective areas. This can be seen in the figure in the lower left of the corresponding images where a very bright area is present. The hue filter generates an edge in the bright area. On the other hand the pseudo-hue filter has problems with the dark areas: the shadows, or in our case the mouth cavity, tend to appear in the filter's results. The filter based on the luminance is failing to 'see' the lips; however, the bright edges seem to be detected very well. Comparing the thresholding neural network approach with the parabolic approach we can conclude that it does a better job in assigning the probability.

C. Defining the Feature Vectors

Here is where the innovation of this paper comes into action. The result of the filter is considered to be the distribution function $I(X,Y)$ of a spatial bivariate distribution. The first observation to be made is that the expected location of this distribution: $[EX,EY]$ approximates the centre of the mouth. Using the mean location we transform the filter's result into polar coordinates using the following formula:

$$J(a,r) = I(EX + r \cos(a), EY + r \sin(a))$$

The algorithm estimates the shape of the mouth by describing the conditional distribution, conditioning on the direction, and then obtain from the distribution function J . For each angle α we, therefore, define the mean and the variance of the conditioned distribution using the following formulas:

$$M(\alpha) = \frac{\int_r J(a,r) r dr}{\int_r J(a,r) dr}, \text{ and}$$

$$\sigma^2(\alpha) = \frac{\int_r J(a,r) (r - M(\alpha))^2 dr}{\int_r J(a,r) dr}$$

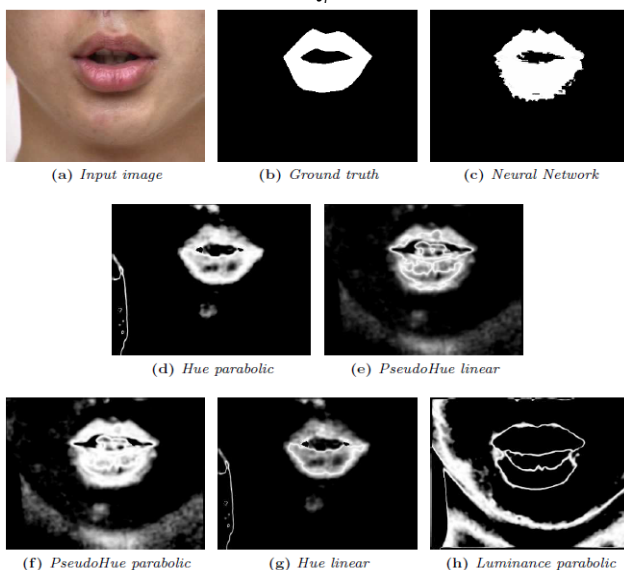


Fig 2 The results of several image filters used to segment the lips pixels.

As an image is discrete rather than continuous, all of the values are obtained from summation rather than integration, so we only operate on estimations of those values, namely $\hat{M}(\alpha)$ and $\hat{\sigma}(\alpha)^2$. For each given angle the conditioning is defined as the circle sector which contains its centre and the vector from the mouth centre which makes the angle α with the horizontal. The rightmost image in Figure 1 shows an example of the estimations of the two values for a number of 18 directions. In this image the values of $\hat{M}(\alpha)$ are shown for each angle and are linked together by a line. It is clearly visible that this polygon passes through the centre of the lips and accurately describes the shape of the mouth. On the other hand, for each angle the perpendicular lines depicted in Figure 1 represent the 95% confidence interval for the conditional distribution, namely it describes the range $[\hat{M}(\alpha) \pm 1.96 * \hat{\sigma}(\alpha)^2]$. So it is obvious that the two sets of values accurately describe the shape of the mouth. The 95% confidence interval clearly describes how thick the lip in that specific direction is, while the means describe the shape of the mouth. We should note that the lips of a wide-stretched mouth appear thinner than those of a closed mouth when related to the overall size of the mouth. The accuracy of the procedure depends on the performance of the image filter used. The difficulties on the filter's side come from the artefacts that can appear on the speaker's face: shadows, areas on the face that have the same colour as the lips, the use of coloured lipstick, etc.; some of these problems have been solved by using an outlier removal technique on the filter's result.

The feature vectors are defined as the vectors containing combined sets of values, computed for a number of angles after the appropriate normalization. Choosing the number of directions is a compromise between accuracy and processing efficiency. The bigger the vectors, the more information on the original distribution they contain but the longer it takes to extract and process them. Also higher dimensionality generally makes it more difficult to train the recognition modules. Wojdel and Rothkrantz indicate that a division of the space into 18 sectors is optimal for obtaining good results [24]. Therefore, we used 18 sectors, shown in Figure 3, to estimate the shape of the mouth, resulting in feature vector of dimension 36.



Fig 3 The 18 feature sectors centred in the centre of the mouth.

D. Visual Validation of the Feature Vectors

Figure 4 shows a sequence of frames with the final annotations and the corresponding features extracted. The feature vectors extracted from the entire recording using the above approach are depicted in Figure 5. The beginning and the end of the utterance are clearly visible as indicated on top of the image. The longer pauses between the words are also

slightly visible. Also visible parts are the round areas located in the variance part which represent the time when the mouth is widely opened.

From this image alone we can conclude that there is significant speech related information in the feature vectors.

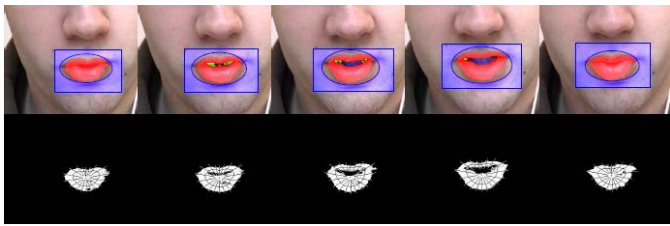


Fig 4 Example of processed frames. The annotation and the extracted features are shown as well

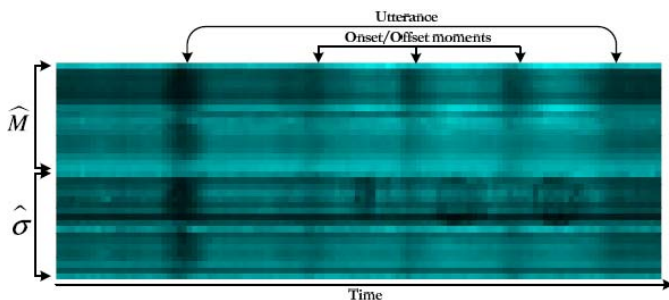


Fig 5 Pairs of $\hat{M}(\alpha)$ and $\hat{\sigma}(\alpha)^2$ vectors extracted from a video sequence. The beginning and ending of the utterance as well as the pauses between the consecutive words are marked by arrows.

Figure 6 shows the samples mean of $\hat{M}(\alpha)$ computed over all utterances that contain the letters A, H and K, respectively. The viseme transcriptions for the three letters are as follows: A = [aa], H = [h aa] and K = [gkx aa]. We can see that in this figure the shapes corresponding to the letters A and K are most of the time overlapping making the classification somewhat difficult. However, as seen in Figure 7, they become more distinguishable when considering the values for $\hat{\sigma}(\alpha)^2$. On the other hand, the situation is almost reversed when analysing in the two images the shapes corresponding to the letters H and K. Therefore, this is an example of a situation when the values of one feature type are not sufficient for discriminating among the three letters. We should remark here that, because in Figure 6 the graph contains the mean of each feature for each angle, we see in fact the mean shape of the mouth during uttering of the corresponding letter. Therefore, this image is not completely reliable since the information about the dynamics of the mouth during speaking are lost due to averaging. Figure 8 shows the standard deviation of $\hat{M}(\alpha)$ which brings back some information about the mouth dynamics.

E. Refinement of the Filter Results: Outlier Removal

As can be seen in the second and the third images in Figure 1, even after reducing the area of interest and even with optimal filtering of the mouth in some cases, the filtered image still contains unwanted artefacts. In these images large artefacts are visible for instance just below the lower lip. This can be the case when there are areas on the speaker's face for which the colour exactly matches the colour of the lips, such as wounds, acne vulgaris or other marks on the face. In order to reduce the impact of such occurrences a process of outlier detection and deletion can be used before the actual feature extraction.

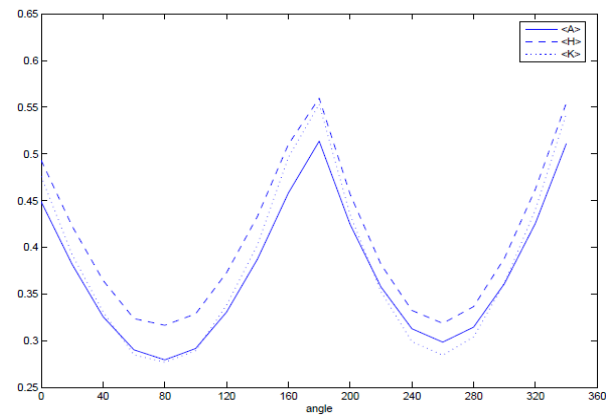


Fig 6 The average of $\hat{M}(\alpha)$ computed over all utterances containing the letters A, H and K, respectively.

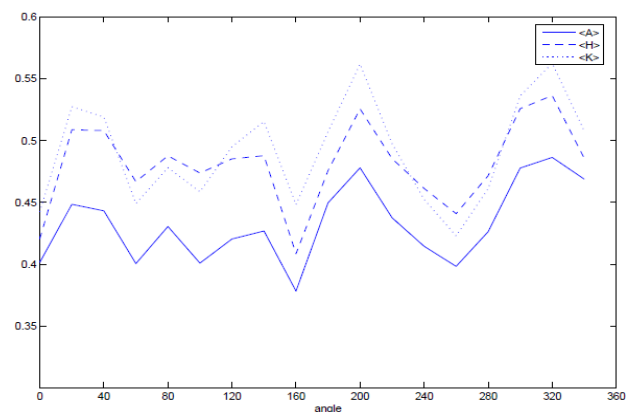


Fig 7 The average of $\hat{\sigma}(\alpha)^2$ computed over all utterances containing the letters A, H and K, respectively.

The blue area shown in the third image of Figure 1 but also in the first row of the images in Figure 4, superimposed on the input image, represents the area which is affected by the outlier deletion process. The algorithm uses the same interpretation of the filter's result, namely as a spatial distribution function, and defines an outlier as an outlier of this distribution. Therefore, an outlier is defined as any location that is further away from the mean by more than a number of standard deviations. The outlier detection process assumes that the image filter, even though it is not perfectly accurate, still assigns the correct probability to a significant number of lip pixels. In Figure 1 we used a rectangular delimitation area, which makes use of a-priori knowledge about the shape of the mouth. Therefore, we applied a different definition of an outlier on the horizontal axis than on the vertical axis. More accurate outlier detection is obtained by defining elliptical delimitation area as shown in Figure 4. The rectangular/elliptical shape refers to the interior edge of the blue area in these images. In Figure 9 we show an example where the necessity of an outlier detection scheme becomes clear. In this figure the images on the left show the results when no outlier detection is used; while on the right we have the results when elliptical outlier detection is used. We can see that the impact on the resulting features can be very large.

F. Cavity, Tongue and Teeth Description

The shape of the lips is not the only visual determinant of a spoken utterance. It was shown that the position of the tongue and the appearance of the teeth while lip reading are important sources of information for the human lip reader too.

However, these elements are not always visible and sometimes their visibility is not only correlated with the spoken utterance but also with the speaking style of the interlocutor. It is essential in the case of lip reading to extract from the visual channel as much information as possible about the utterance, especially given the fact that compared with the audio modality it is agreed that the visual speech inherently provides less information. We propose that this type of information should be used to enrich the feature vectors in any case. Tracking of the teeth is easier than tracking the tongue. The teeth are much brighter than the rest of the face and can therefore be located using a simple filtering of the image intensity. The visibility and the position of the tongue cannot be determined as easily as in the case of the teeth, because the colour of the tongue is almost indistinguishable from the colour of the lips. We can, however, easily determine the amount of mouth cavity that is not obscured by the tongue. While teeth are distinctly bright, the whole area of the mouth behind the tongue is usually darker than the rest of the face. So we can apply an intensity based thresholding filter for both cases. The teeth and cavity areas are both highlighted in Figures 1 and 4, the cavity in green and the teeth in dark blue. In order to describe the appearance of these two elements into quantitative data we used the total area of the highlighted region and the position of its centre of gravity relative to the centre of the mouth. Therefore, the feature vectors were enlarged with 6 more entries for every frame.

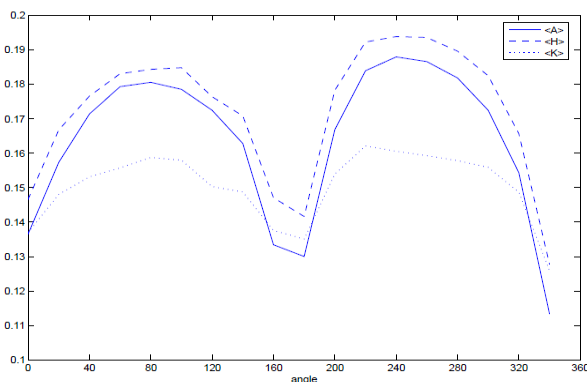


Fig. 8 The standard deviation of $\hat{M}(\alpha)$ computed over all utterances containing the letters A, H and K, respectively.

IV. RESULTS OF EXPERIMENTS

We used the algorithms described in this paper to process our data corpus. For each recording a file was produced containing one feature vector for each video frame in the recording. Each vector had 42 components, 2×18 shape features, namely for 18 angles the mean and variance of the conditional distributions and 6 intensity features describing the presence of the teeth and the cavity. We trained a lip reader based on the HMM approach for each recognition task, namely for connected digits, connected letters, grammar based utterances, random sentences and the complete corpus. The best result was achieved for the digit recognition task from which we obtained 43.24% word recognition rate and 33.59% accuracy rate. There are several ways to improve the basic performance of the system. We considered the inclusion of the first and second derivatives of the feature vectors. This meant that the new feature vectors would have 126 entries. In the first group of settings each state of the HMMs used for inference contains a 42-dimensional Gaussian, while in the second group of settings this changes to a 126-dimensional Gaussian. As a result, the number of parameters that need to

be computed during the training stage increases many folds with the new settings. With enough data in the training set to train all the new parameters, the performance of the new recognisers as expected to increase considerably.

We found that for the best case, namely the digit recognition, the word recognition rate was 54.05% and the word accuracy was 43.24%. In the case of the letter string recognition task the increase was smaller, from 29.48% word recognition rate to 34.33%. We think that this can be explained by the fact that the size of the corpus used for the letter recognition task was smaller than in the case of the digit recognition task. In each state the distribution over the observed features is approximated using a Gaussian distribution. However, the real distribution of the observations is rarely Gaussian. Therefore, in order to make the approximation better, a mixture of Gaussian distributions is used. Since the number of Gaussian distributions best to approximate the real distribution is not known in advance, a trial and error approach is usually used. In our case this approach proved to be very useful since the performance of the recognizer was substantially increased. For instance in the case of the digit recognition task, when only the static features were used we obtained a word recognition rate at 58.69% and word accuracy at 49.81%.

This result was obtained from 32 Gaussian mixtures. In the same experiments, but when the dynamic features were added, the results increased as well to a maximum of 74.52% word recognition rate and 59.46% word accuracy with 25 Gaussian mixtures used. All the results above were obtained from systems that considered isolated visemes as building blocks. This means that influence is from neighbouring visemes is not considered. However, in speech recognition there is a good practice to use the left and the right viseme to build a local context. This is only possible when there is sufficient data in the training corpus so that the number of unseen context combinations is very low. In the case of continuous speech recognition tasks this is very difficult to obtain. However, we were able to use this approach for the digit and letter recognition tasks with great success. For instance, when the full feature vectors combine the static and the dynamic features, the increase in word recognition rate is from 54.05% to 69.88%. When combining all the above approaches we obtained for the digit recognition task the best results when using a 28 Gaussian mixture, namely a word recognition rate at 89.96% and a word accuracy at 76.83%. Figure 10 shows the graph of word recognition rate and words accuracy as a function of the number of Gaussian mixtures for the case of the CD task where all the above tuning was used. It is clearly visible in this case that having sufficient data for training using a Gaussian mixture to approximate the state distribution has a great impact.

The results for the other more complex recognition tasks, such as random sentences, grammar based utterances, continuous speech were, as expected, are less impressive. However, they are very promising because they are great improvements over the previous results.

For instance in the case of the GU recognition task the best result obtained was WRR 48.36% and Acc 10.33%, while for the ALL recognition task the best result obtained was as small as WRR 15.94% and Acc -8.96%. Due to the difference in the complexity of the recognition tasks, a decrease in performance was expected. However, especially

in the latter case we investigated more in depth to find the exact reasons.

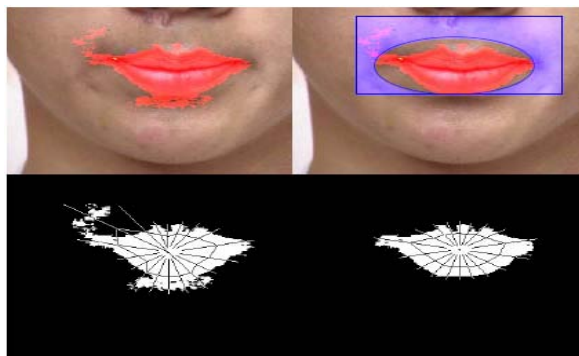


FIG. 9 Outlier detection for improving the estimation of the lip geometry.

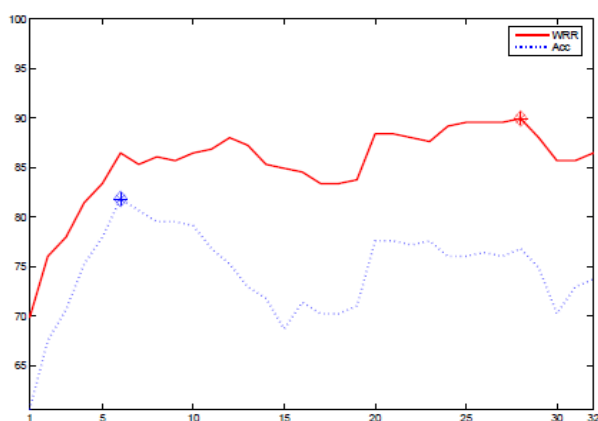


FIG. 10 The WRR and Acc results for CD recognition task as a function of the number of mixtures. The x axis gives the number of mixtures and the y axis shows the results obtained. The recognition system consisted of context aware HMMs trained on 126-dimensional feature vectors.

This was also because the results show a great difference between the word recognition rate values and the accuracy values. This difference appears because of a very high insertion rate. This is a big problem in the lip reading domain because anytime the mouth is moving the system may interpret it as speaking. This is not the case for speech recognition when the silence is very well defined. Therefore, a better definition of the silence models is needed. The insertion error is, however, not the only question to ask here. When we looked at the detailed results, we found that substitution error level was even higher. For instance in the GU task reported above the substitution rate was 39.93% while the insertion level was 38.04% and in the ALL task the substitution rate was 74.02% while the insertion rate was 24.90%. This picture shows that the systems are not very well trained in general; therefore, showing the amount of data samples in the training set is well under the optimum limit. In order to further investigate this, we build recognition systems for which we removed the language modelling layer, which means that the results are obtained in term of viseme strings rather than in terms of word strings. For instance in the case of the GU task the results were WRR 56.89% and Acc 15.30%, substitution rate 25.07% and insertion rate 41.59%. There are two things to remark here. Firstly, the recognition performance has increased. Secondly, and more importantly while the substitution has decreased the insertion rate has increased considerably. These results explain the low recognition task

but it does so in agreement with our expectation that most errors are made in the non speaking part of the utterances.

An increase in performance was observed in the case of the ALL recognition task as well: WRR 39.23% and Acc: 20.56%. In order to get a better idea about the recognition results we investigated the confusion matrices. It is possible to visualize the viseme confusion matrix in all cases when the results are given in terms of viseme strings. However, when the results are given in terms of word strings, visualising the confusion matrix is feasible only in the case of digit strings and letter strings. In Figure 11 the images a) and b) show the confusion matrix obtained by the most successful recognisers. In the case of digit recognition the confusion matrix is almost perfectly diagonal. For the alphabet letters this is not the case anymore. We notice that the most confused digit is the digit < 1 > ([iee gkx]) and is often confused with digit < 9 > ([gkx iee gkx at]). On the other side the letter < A > ([aa]) is confused with < H > ([h aa]), the letter < C > ([sz iee]) is confused with < D > ([td iee]), the letter < G > ([gkx iee]) with < D > ([td iee]), the letter < N > ([eeh gkx]) with < L > ([eeh l]), etc.

Exactly the same pattern appears irrespective with the number of mixtures we used. The images c) and d) from the same figure show the mean confusion matrices for each case. It can be seen that when a large part of the word transcription is similar, the confusion increases. It should be noted that the confusion matrix only shows the substitutions and in some small extent gives some insights about the deletions. The confusion matrix has some empty rows and columns. These elements correspond to the letters that due to the viseme definition have similar transcription in the viseme space.

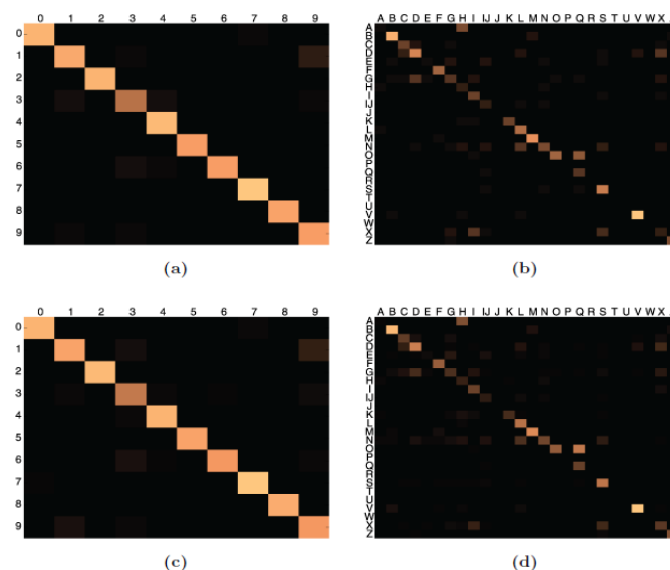


Fig. 11 The confusion matrices obtained by the best systems in the CD and CL tasks, respectively. a) the confusion matrix for CD task in the best case. b) the confusion matrix for CL task in the best case. c) the mean, over the mixture number, confusion matrix for the CD task. d) the mean, over the mixture number, confusion matrix for the CL task.

We also investigated the results in an N-Best approach where the first 5 possible outcomes were considered. This approach gives the possibility to post process the results in order to choose the most probable outcome. However, the increase in the best result was rather marginal.

Even though it is not a common practice in speech recognition, but because in our case we still have a relatively small corpus compared with a regular corpus used for speech

recognition, we analysed the results in a 10-fold validation experiment. In a 10-fold validation experiment the data is divided into 10 folders and each time 9 folders are used for training and 1 for testing. This means that we trained on 90% of the data and tested on 10% of the data in the corpus. This approach is meant to increase the certainty in the results, namely, one expects to have a low variance in the results obtained on different data. This is exactly what we found in our experiments. For instance in the case of the CD task we found the mean word recognition rate over the 10 folds was 88.82% with a standard deviation of 1.48%. The maximum word recognition rate was 91.39%. Similarly, the mean accuracy rate found was closer to the one obtained in the first experiment, namely 74.50% with a standard deviation of 3.73%. The maximum accuracy obtained was 79.43%. In all the experiments we noticed a very large insertion rate which agrees with the previous findings. The best results are summarized in Table 1.

TABLE I
THE BEST RESULTS OBTAINED BASED ON THE SLGE FEATURE EXTRACTION METHOD

Task	WRR	Acc
CD	91.39%	79.43%
CL	56.34%	17.16%
GU	56.89%	15.30%
All	39.23%	20.56%

V. CONCLUSIONS

In this chapter we investigated the use of a statistical approach for estimating the shape of the lips for lip reading. We presented two methods for improving the process of lip geometry estimation, namely, region of interest detection and outlier detection. These improvements make the method more robust to the changes in performance of the colour filter used. By using an object detection algorithm to find an accurate bounding rectangle around the mouth we remove much of the face areas on which the colour filter is prone to make errors. Even further, the outlier detection approach produces smoother and more accurate mouth shapes. From the point of view of the processing complexity, we found that this approach is very fast, and could be successfully deployed in real time applications. However, we should stress here that the development of a suitable colour space and accompanying filter would make this method more universally applicable.

The results obtained based on this approach show great improvements over previous results. In the case of the simpler tasks, like digit strings and letter string recognition we achieved results comparable to the previous results. However, in the case of the more complex utterances we found great improvements. This result is a clear indication that this method can be successful for more complex systems. However, in order to achieve good performance we need more data and data of better quality (i.e. we have shown in the previous chapters that the corpus we used was specially built for the current research, and that it is considerably larger than the previous data corpus).

REFERENCES

- [1] J. C. Wojdel, and L. J. M. Rothkrantz, "Obtaining Person-independent Feature Space for Lip reading", in Proceedings of AVSP, 2001 *Visually based speech onset/offset detection*.
- [2] L. J. M. Rothkrantz, and J. C. Wojdel, "Fusing data streams in continuous audio-visual speech recognition", in Proceedings of Text, Speech and Dialogues, 2005, Springer.
- [3] McGurk, and J. Macdonald, "Hearing lips and seeing voices", *Nature*, vol. 264, pp. 746-748, December, 1976.
- [4] J. J. Williams, J. C. Rutledge, and A. K. Katsaggelos, "Frame rate and viseme analysis for multimedia applications to assist speech reading". *Journal of VLSI Signal Processing*, vol. 20:pp. 7-23, 1998.
- [5] J. N. Buchan, M. Pare, and K. G. Munhall, "Spatial statistics of gaze fixations during dynamic face processing", *Social Neuroscience*, vol. 2(1), pp. 1-13, 2007.
- [6] S. Hilder, R. Harvey, and B. J. Theobald, "Comparison of human and machine-based lip-reading", in B. J. Theobald and R. W. Harvey, editors, AVSP 2009, pp. 86-89. Norwich, September.
- [7] E. Petajan, B. Bischoff and D. Bodoff, "An improved automatic lip reading system to enhance speech recognition", In CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 19-25. ACM Press, New York, NY, USA.
- [8] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lip reading", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 198-213, 2002.
- [9] N. Li, S. Dettmer, and M. Shah, M. "Lip reading using eigen sequences", in Proc. International Workshop on Automatic Face- and Gesture-Recognition, pp.30-34. Zurich, Switzerland, 1995.
- [10] N. Li, S. Dettmer, and M. Shah, "Visually recognizing speech using eigensequences, Motion-based recognition", vol. 1, pp. 345-37, 19971
- [11] J. Luetin, N. A. Thacker, and S. W. Beet, "Statistical lip modeling for visual speech recognition", in Proceedings of the 8th European Signal Processing Conference (EUSIPCO96).
- [12] J. Luetin, and N. A. Thacker, "Speech-reading using probabilistic models", *Computer Vision and Image Understanding*, vol. 65(2), pp. 163-178, 1997.
- [13] I. Arsic, and J. P. Thiran, "Mutual information eigenlips for audiovisual speech recognition", in 14th European Signal Processing Conference (EUSIPCO, 2006).
- [14] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lip reading", In Proc. IEEE International Conference on Image Processing, vol. 1, 1998.
- [15] S. Dupont, and J. Luetin, "Audio-visual speech modeling for continuous speech recognition", in *IEEE Transactions On Multimedia*, vol. 2. September, 2000.
- [16] C. Wojdel, Automatic Lip reading in the Dutch Language, Ph.D. thesis, Delft University of Technology, November, 2003.
- [17] G. Potamianos, C. Neti, J. Luetin, and I. Matthews, "Audio-visual automatic speech recognition: An overview", *Issues in Visual and Audio-Visual Speech Processing*, 2004.
- [18] J. F. G. Perez, A. F. Frangi, E. L. Solano, and K. Lukas, "Lip reading for robust speech recognition on embedded devices", in *Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 473-476, 2005.
- [19] P. Lucey, and G. Potamianos, "Lip reading using profile versus frontal views", in *IEEE Multimedia Signal Processing Workshop*, pp. 24-28, 2006.
- [20] V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition", *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1274-1288, 2002.
- [21] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speech reading with applications to human-computer interfaces", *EURASIP Journal Applied Signal Process*, vol. 2002(1), pp. 1228-1247.
- [22] K. Kumar, T. Chen and R. M. Stern, "Profile view lip reading", in Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP, vol. 4, pp. 429-432, 2007.
- [23] P. Viola, and M. Jones, M. "Robust Real-time Object Detection", in Second International Workshop On Statistical And Computational Theories Of Vision Modelling, Learning, Computing, And Sampling, Vancouver, Canada, July, 2001.
- [24] G. Chitu, L. J. M. Rothkrantz, P. Wiggers, and J. C. Wojdel, "Comparison between different feature extraction techniques for audio-visual speech recognition", in *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pages 7-20, Springer, March, 2007.